An Experimental Study of Old and New Depth Measures

John Hugg, Eynat Rafalin, Kathryn Seyboth, Diane Souvaine
Tufts University

Department of Computer Science, Tufts University, Medford MA 02155
Partially supported by NSF grant CCF-0431027

Overview

1 Introduction to Data Depth

"What is a Depth Measure?"

2 The Depth Explorer Sandbox

A software tool to compare Depth Measures

3 Experimental Evaluation of New and Existing Depth Measures Using Depth Explorer

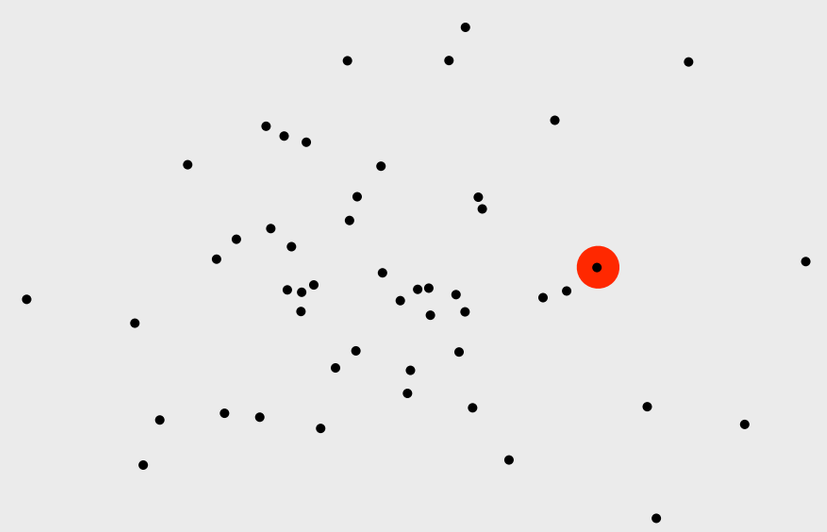
What is Data Depth?

Data depth is a statistical method to analyze data

Assigns a numeric value to a point relative to its centrality in a data set.

Examples: Half-space depth, convex-hull peeling depth, L1 depth

Example:



Depth of this point = ?

The *Halfspace Depth* of a point x relative to a set S is the minimum number of points of S lying in any closed halfspace passing through x.

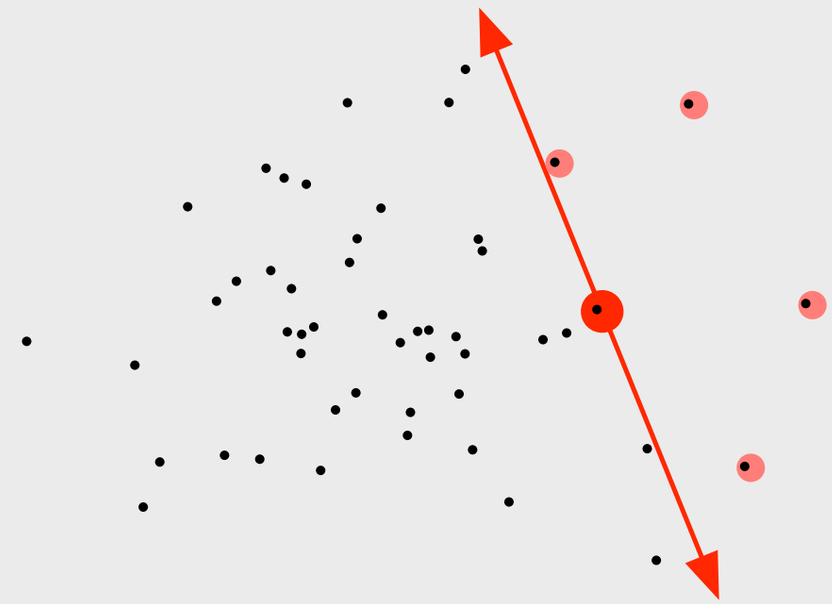
What is Data Depth?

Data depth is a statistical method to analyze data

Assigns a numeric value to a point relative to its centrality in a data set.

Examples: Half-space depth, convex-hull peeling depth, L1 depth

Example:



Depth of this point = 4

The *Halfspace Depth* of a point x relative to a set S is the minimum number of points of S lying in any closed halfspace passing through x.

What is Data Depth?

The Body of Research is the Field is Growing...

Statisticians

Tukey '74

Eddy '82

Liu '90

Donaho & Gasko '92

Liu et al '99

Zuo & Serfling '00

More...

Computer Scientists

Chan '95

Matousek '91

More...

What is Data Depth?

- Non-Parametric

  Infer Information From the Actual Data with an Unknown Distribution (Without Imposing an Underlying Distribution) In Contrast to Traditional Statistics

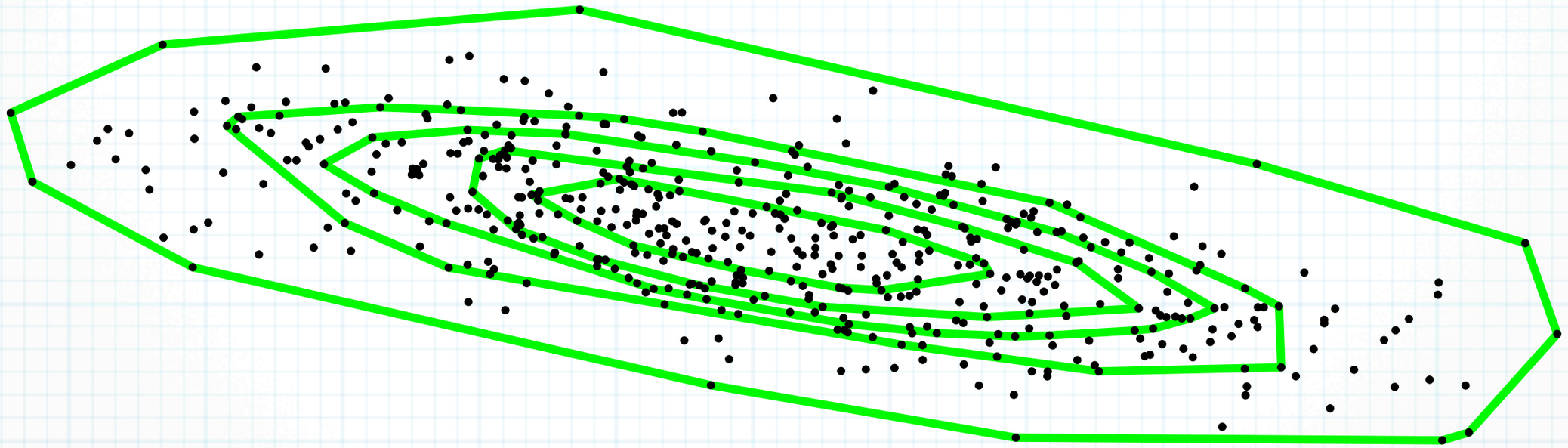
- Center-Outward Ordering of Points
- Robust / Not Sensitive to Outliers

  Outliers Minimally Affect the Depth of Interior Points

- Applies to Multivariate Data Sets

  We Can Use Data Depth To Analyze High-D Data

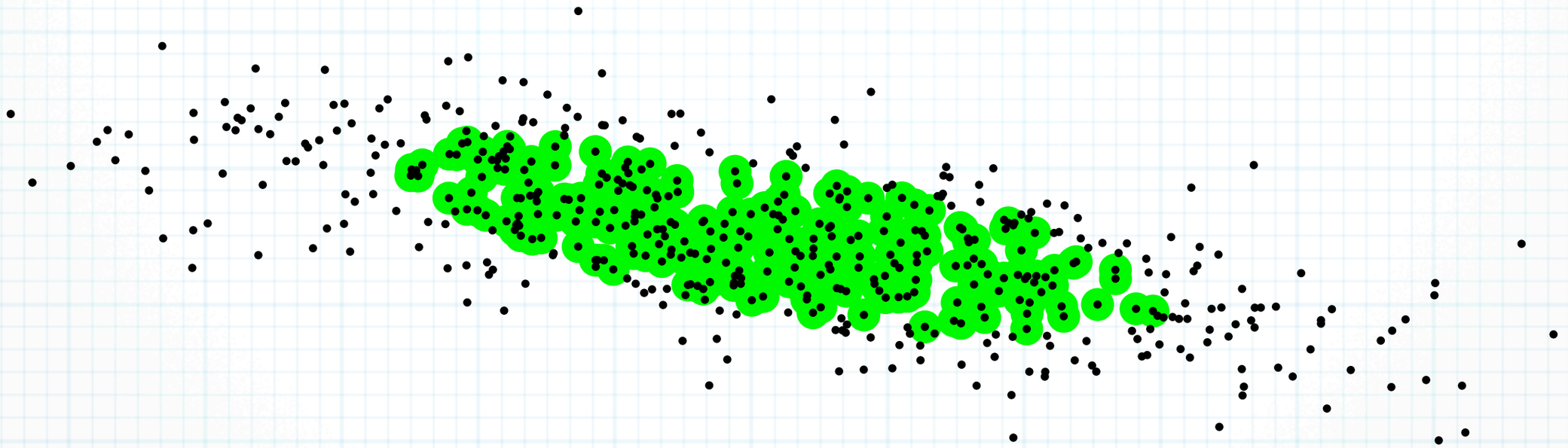
Depth Contours



- Depth Contours are nested contours that enclose regions of increasing depth
- Depth contours help visualize the shape of the data

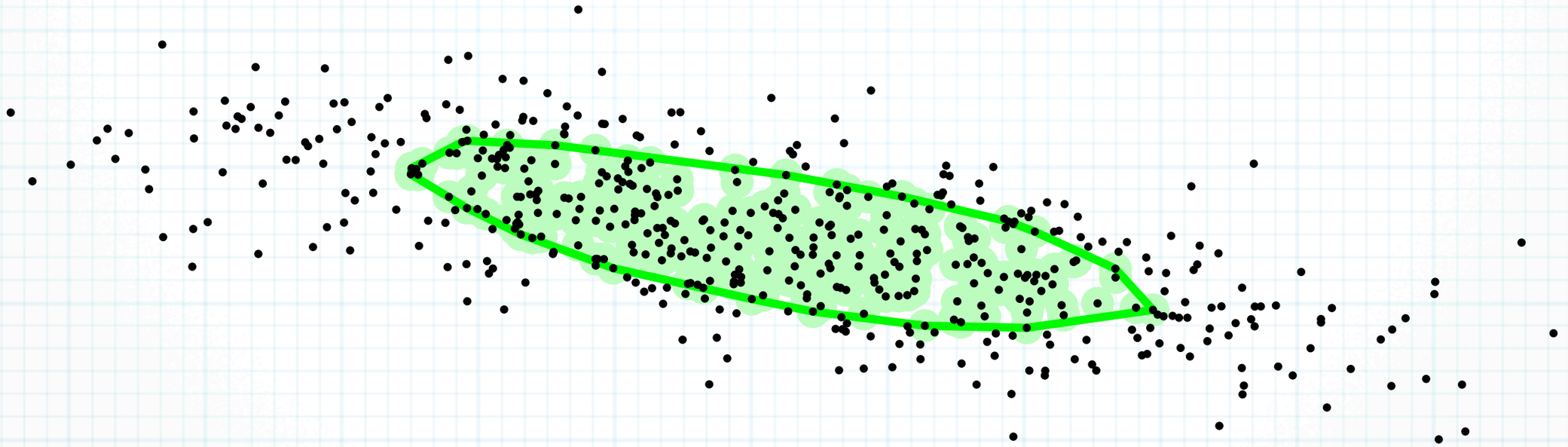
  They can provide a "Topological Map"
- Can determine outliers

Depth Contours



- The region enclosed by the contour of depth t is the set of points such that $D(x) \geq t$
- For well behaved depth function the contours can be approximated using the convex hull of the point of depth t *[Liu 2003]*

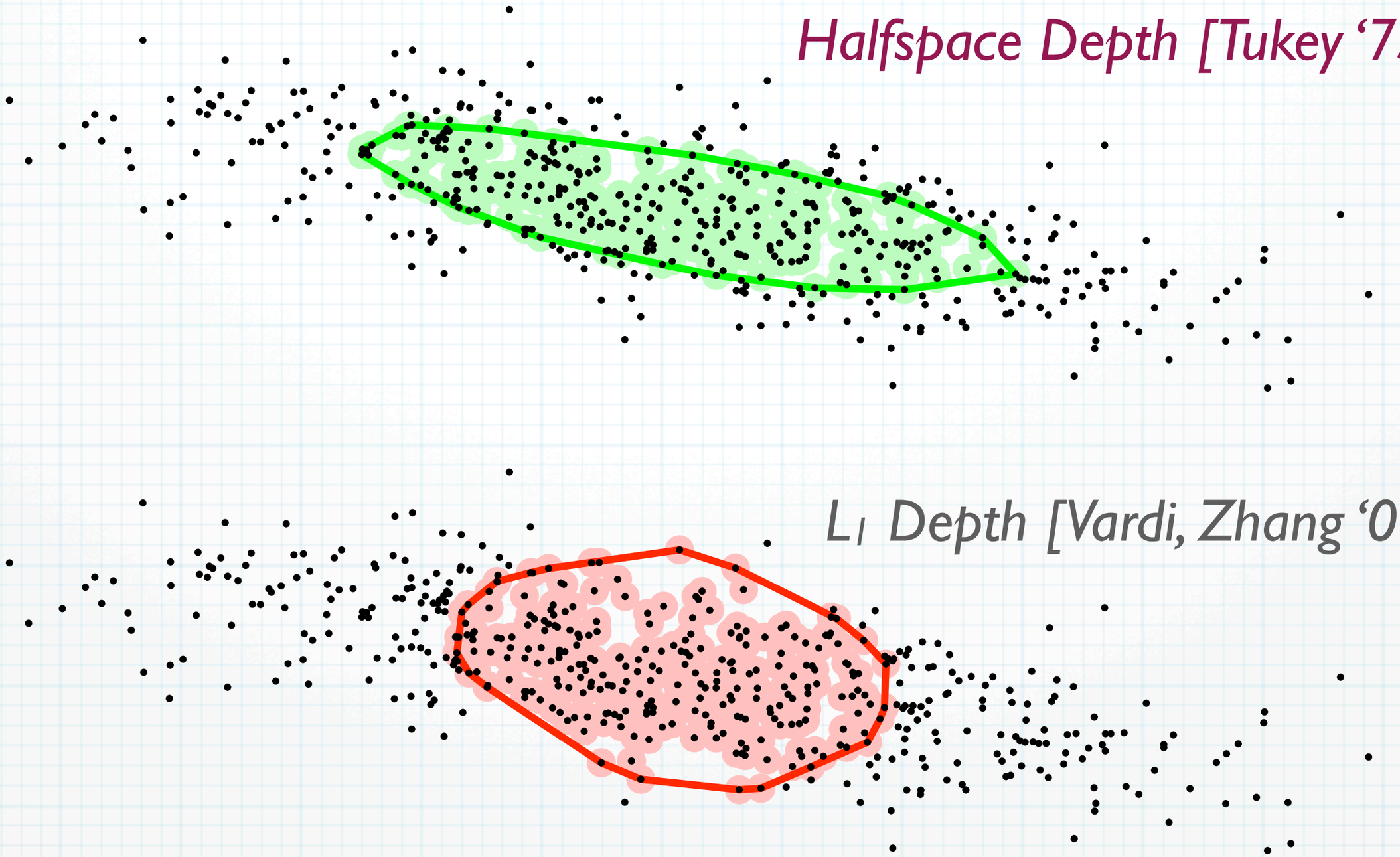
Depth Contours



- The region enclosed by the contour of depth t is the set of points such that $D(x) \geq t$
- For well behaved depth function the contours can be approximated using the convex hull of the point os depth t

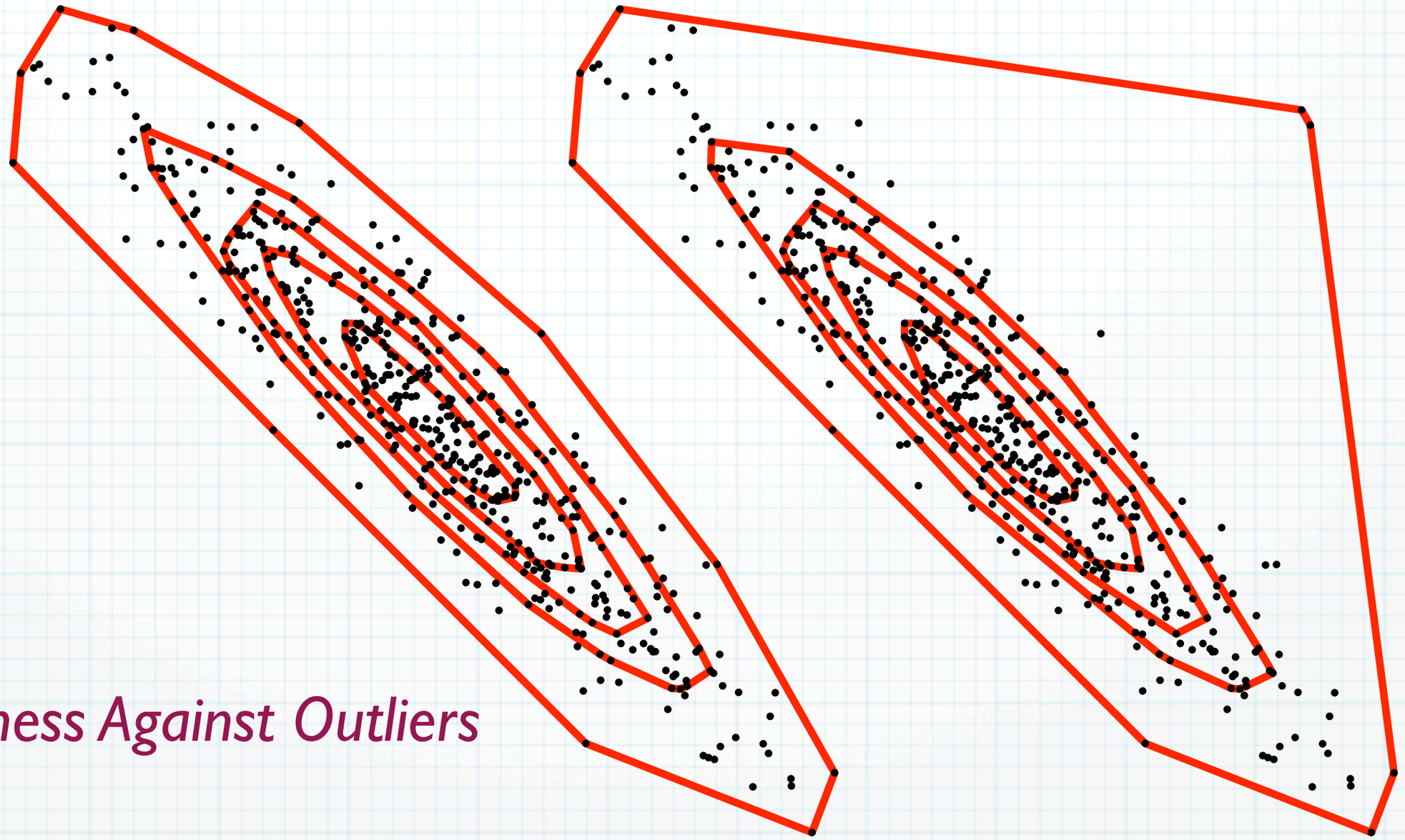
Contour Examples

Halfspace Depth [Tukey '75]



L_1 Depth [Vardi, Zhang '00]

Contour Examples



Robustness Against Outliers

So What's the Problem?

- Popular Depth Measures in 2D don't extend well to High Dimensions

  Many Require $O(n^d)$ Time to Compute

- Depth Measures that are fast in High-D often have drawbacks or untested statistical behavior

- Without better High-D depth measures, lots of data sets can't use depth

  e.g. Bioinformatics

Depth Explorer

What is Depth Explorer?

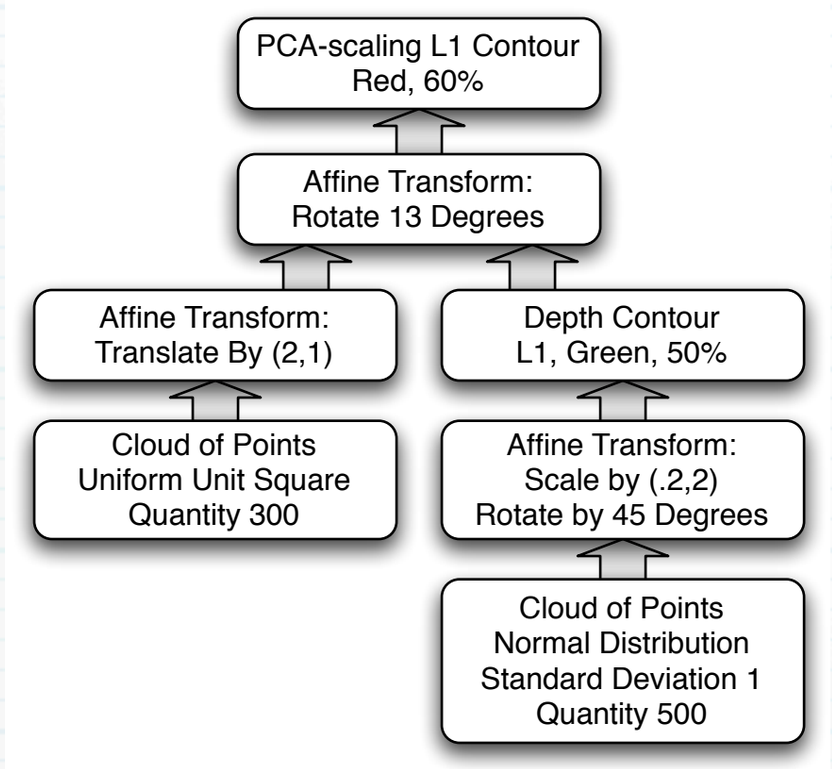
- Graphical, interactive program.
  Provides Fast Feedback

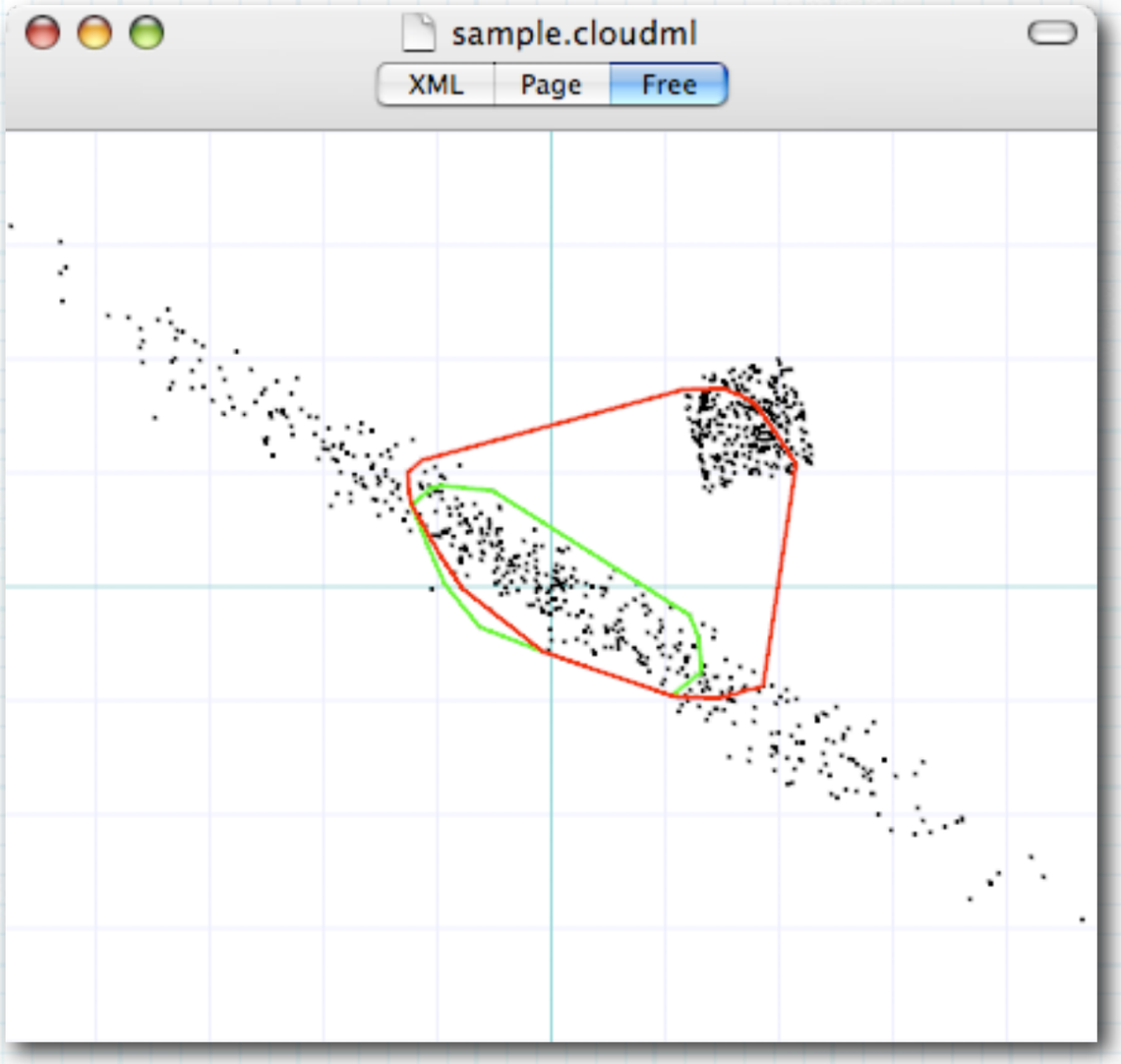
- Facilitates the comparison of Depth Measures
  7 Today - More to Come

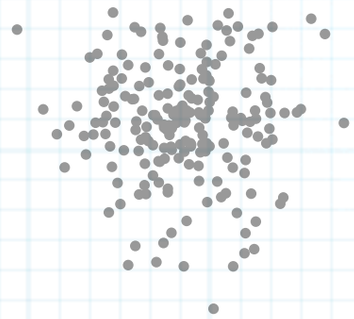
- Load/Save/Print/Help/Etc...
  General Application Functionality

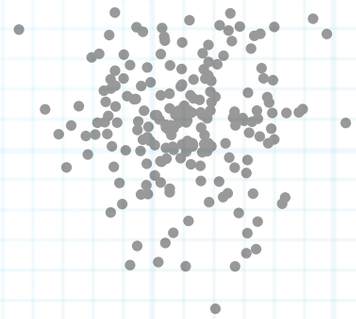
How it works: The Render Tree

```
<canvas width='6' height='6' margins='.1' minrange='6'>
    <pcabag color='1,0,0' size='.6'>
        <transform angle='13'>
            <transform xtrans='2' ytrans='1'>
                <cloud type='uniform' points='300'/>
            </transform>
            <bag type='l1' color='0,1,0' size='.5'>
                <transform xscale='.2' yscale='2' angle='45'>
                    <cloud type='normal' points='500'/>
                </transform>
            </bag>
        </transform>
    </pcabag>
</canvas>
```

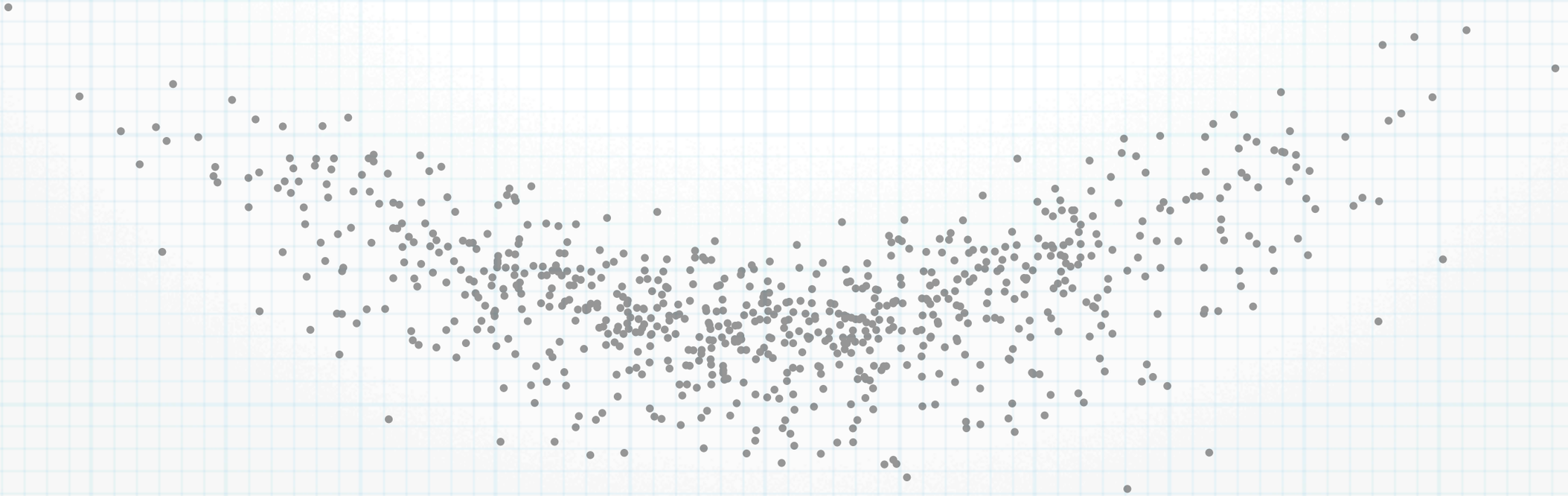




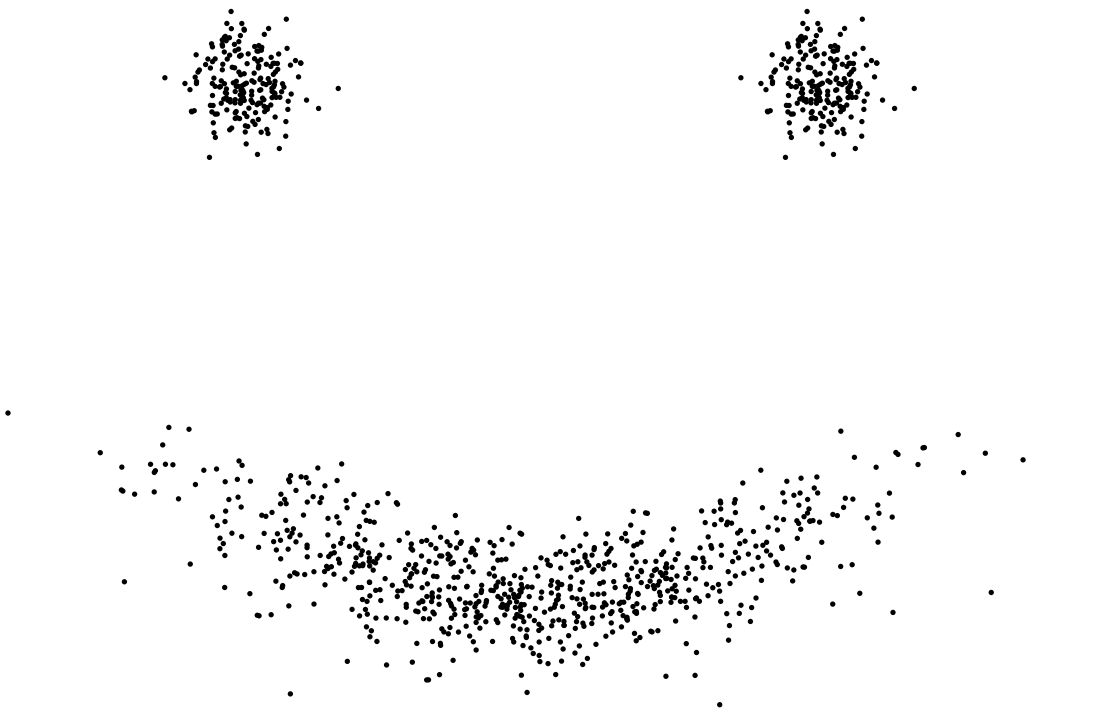




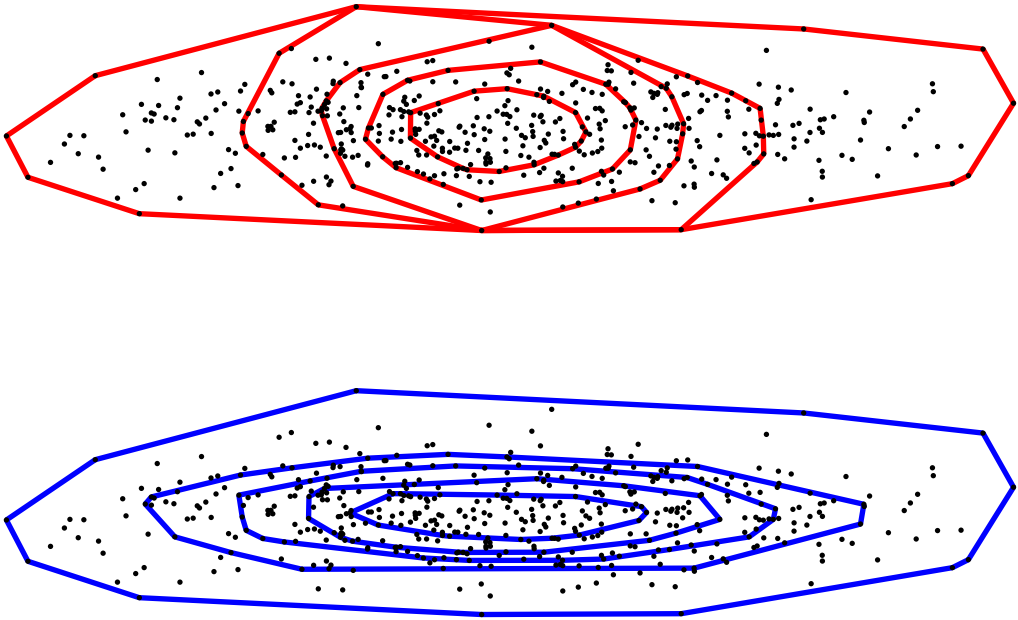
DEMO



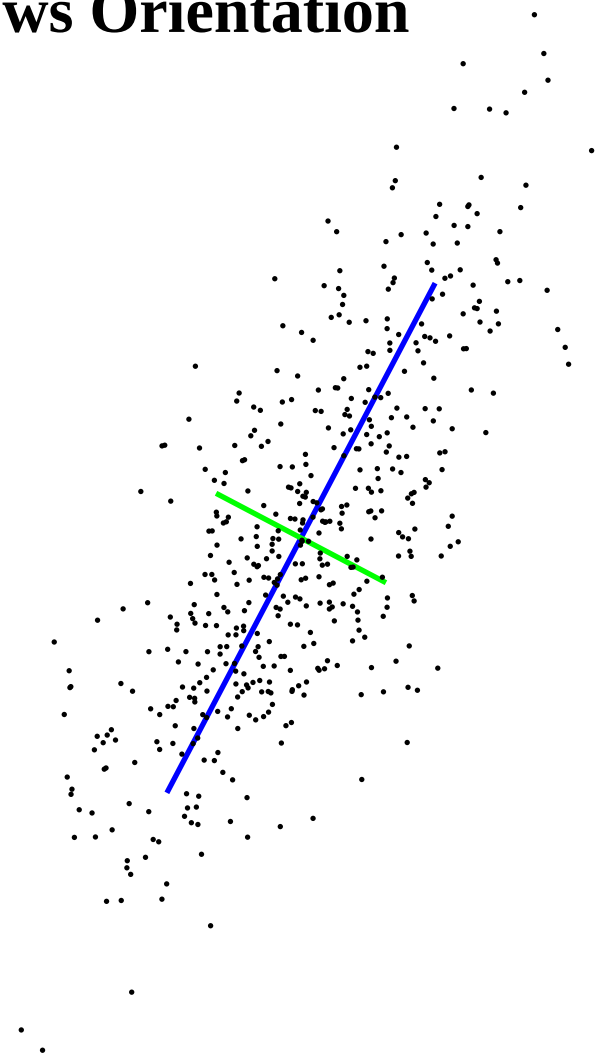
Smily Face



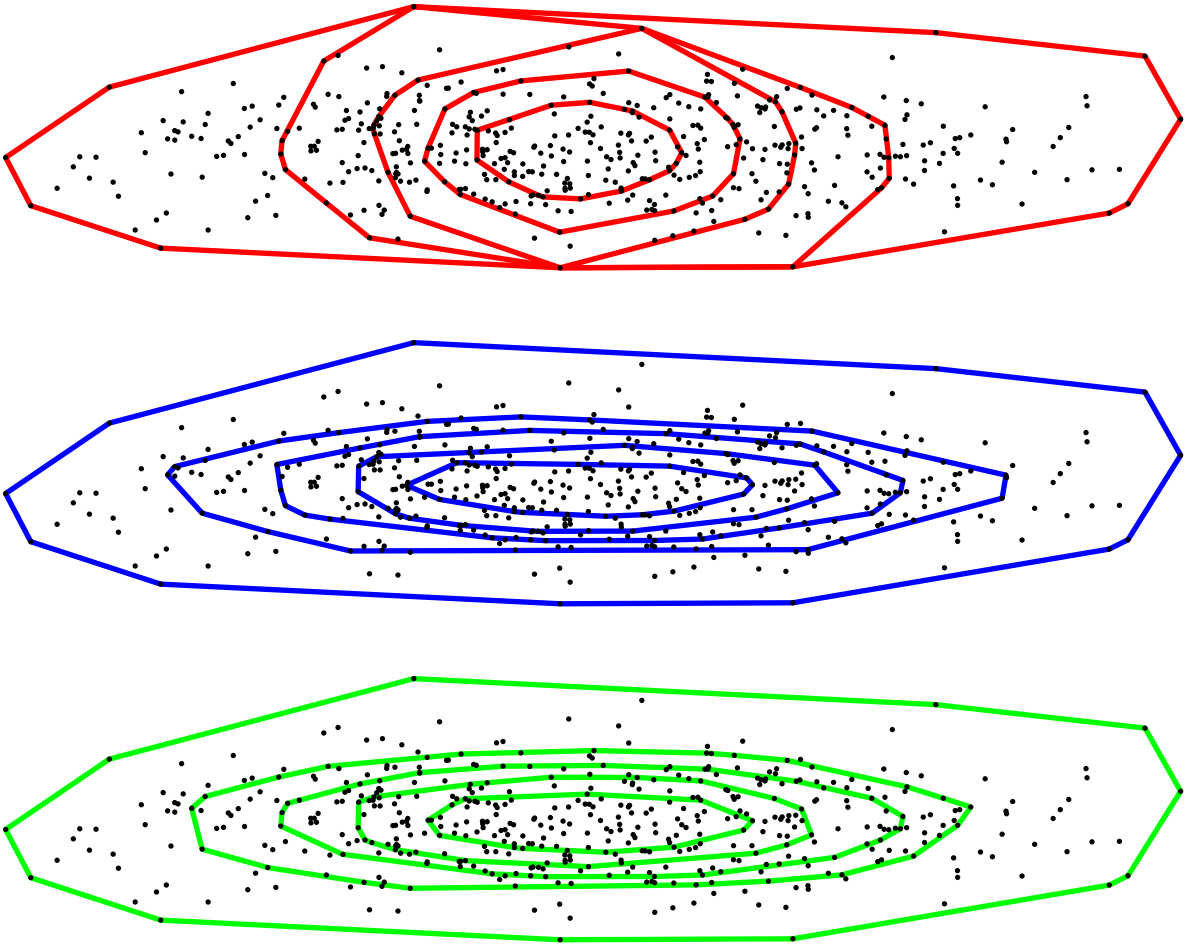
Flaws with the L1 Depth Measure



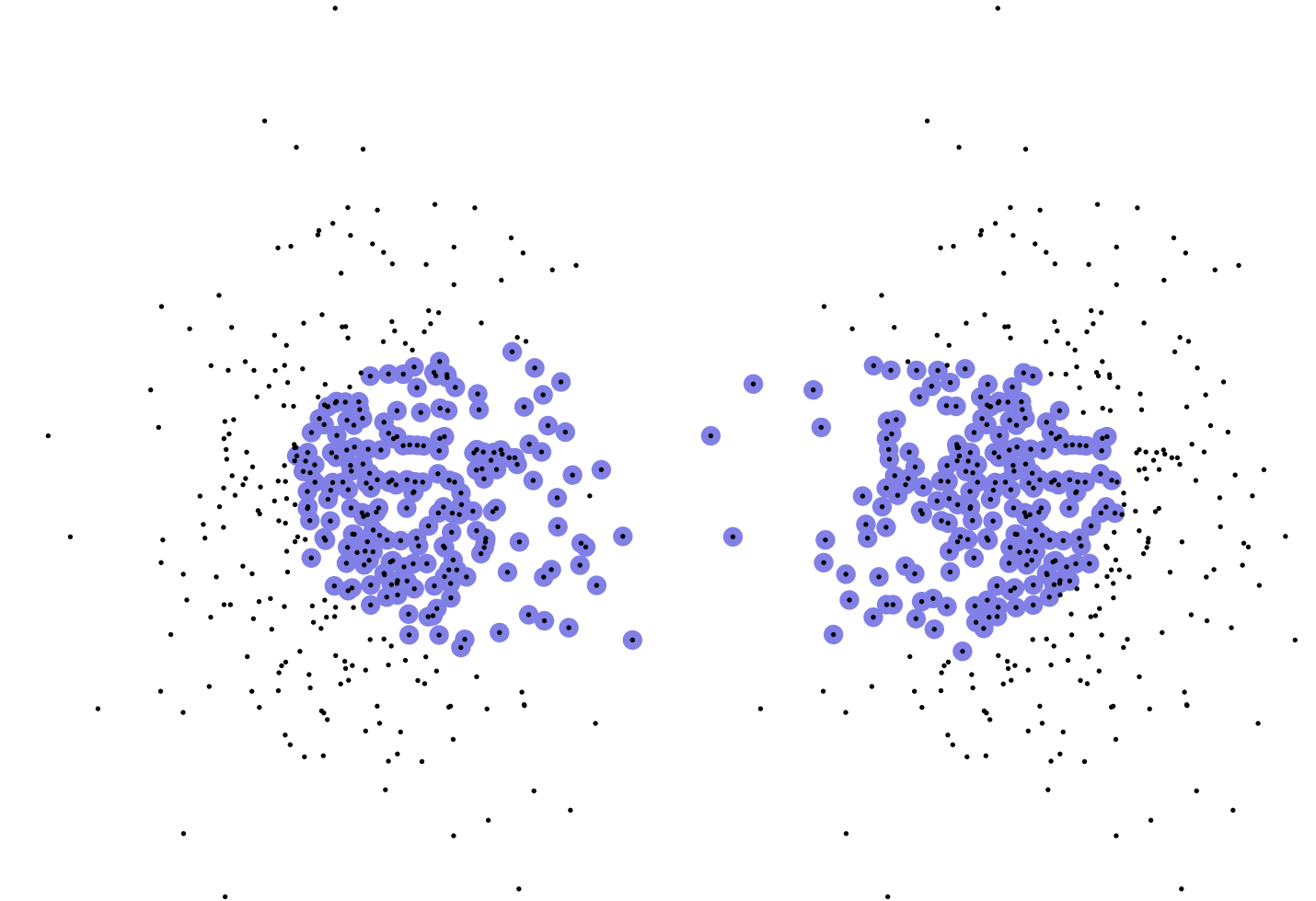
PCA Shows Orientation



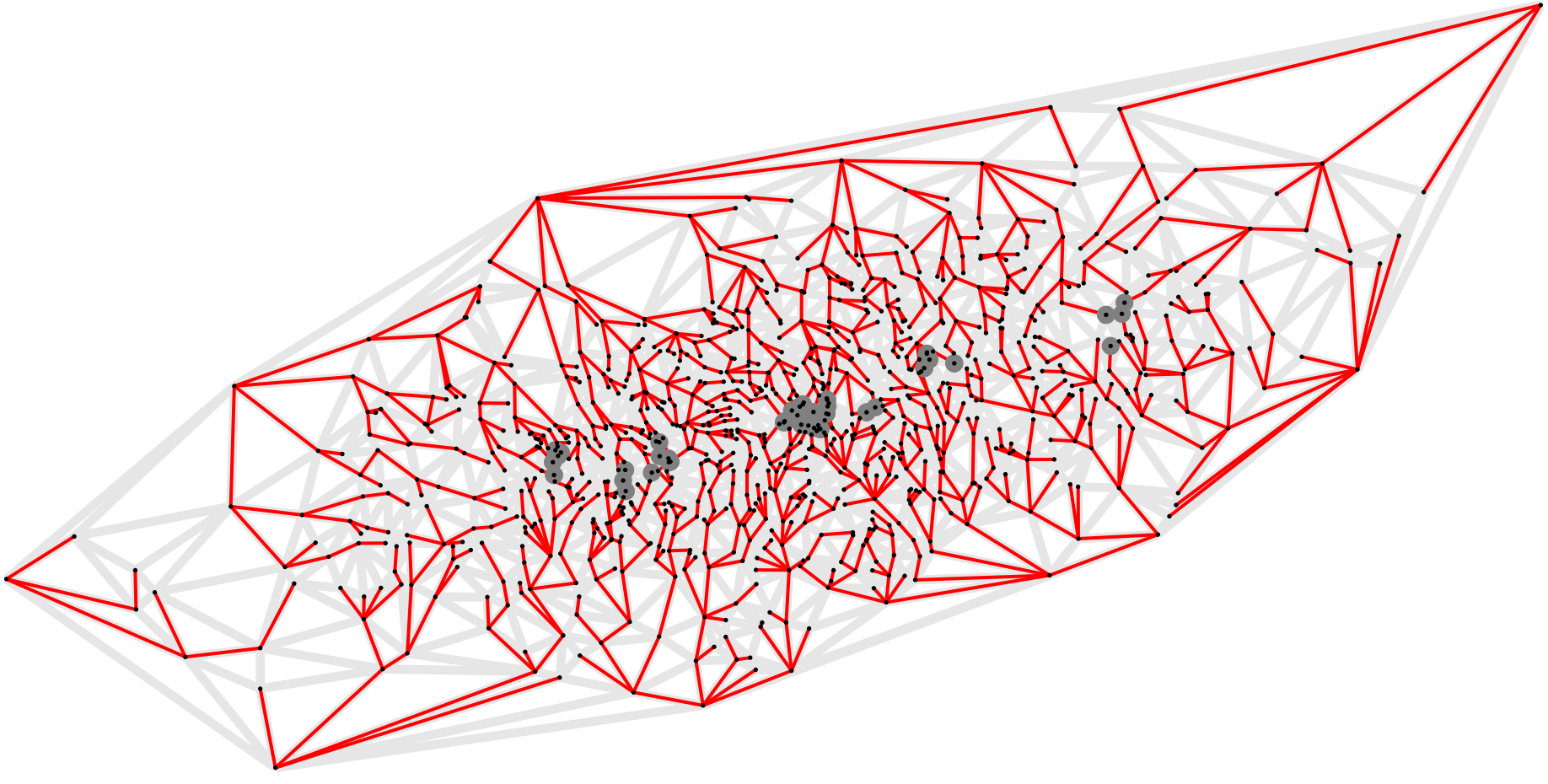
L1 Depth vs. New PCA Scaling L1 Depth



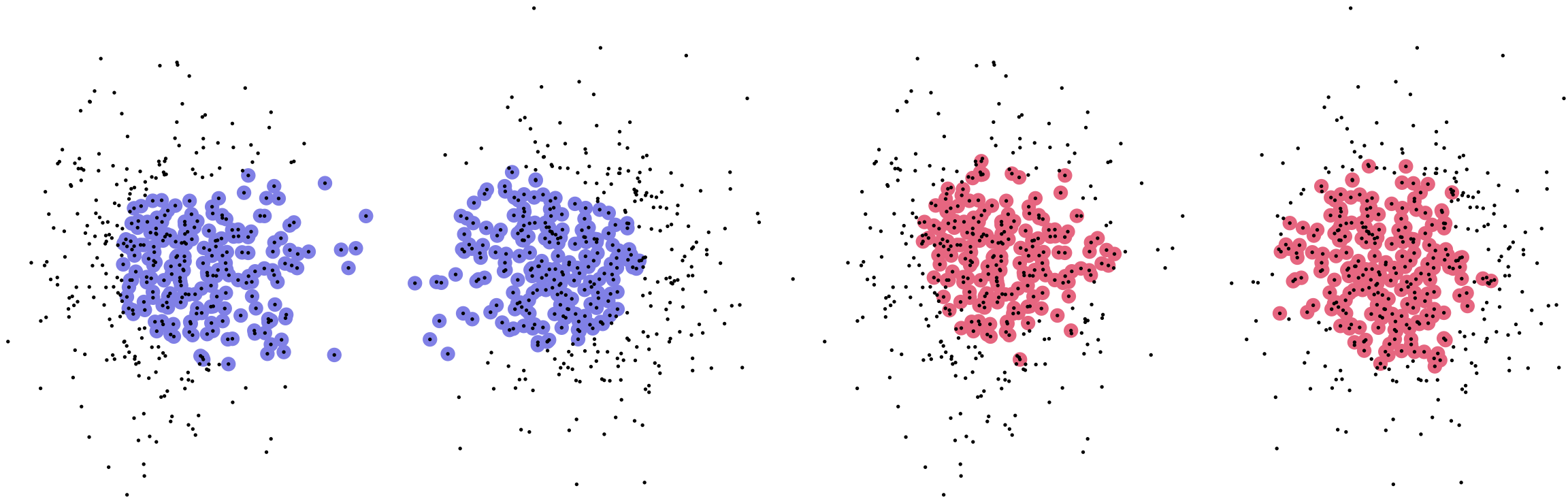
Bimodal Data Cloud



Proximity Trees Generating Depth Values



Proximity Depth Captures Bimodality



Summary

- Data Depth is a powerful tool that can augment traditional statistics.
- While current Data Depth concepts have proven very useful for some problems, more work is needed to apply Data Depth to others.
- In a young field like Data Depth, software like Depth Explorer is a valuable aid in evaluating the behavior of unproven depth measures on many different data sets.

Depth Explorer in 2006

<u>Version 1.0 (Today)</u>

- Available free on the web for Mac OS X 10.4
- Supports 7 Depth Measures
  Convex Hull Peeling, Halfspace, L1, PCA-L1, Proximity (3 Kinds)
- Can Highlight Points or Draw Contours to Indicate Depth
- Supports XML Saving, PDF-Export, Online Help and many other Desktop Application Features

Depth Explorer in 2006

More Releases Throughout 2006

- Improved Documentation & Tutorial
- More Modular Code to Facilitate Extensions
- Higher Dimensional Statistics

Depth Explorer in 2006

How the Tufts Geometry Group Will Use DE in 2006

- Exploring Data-Depth-Based Robust Clustering
  Depth Explorer Will Color Points to Show Clusterings

- Visualizing the Multivariate Box Plot, the Bag Plot
  The Bag Plot is Useful for Detecting Outliers

- Developing Depth Measures that Give Meaningful Results on Non-Convex Data Sets
  Currently, Depth is Difficult to Apply to Banana Shaped Data

For More Info:

http://www.cs.tufts.edu/r/geometry/